UNITED STATES DISTRICT COURT
FOR THE DISTRICT OF MICHIGAN

ONUR BASER,

Plaintiff,

v. Civil Action No. 13-12591

UNITED STATES DEPARTMENT OF
VETERANS AFFAIRS

Defendant.

______________________________________/

SUPPLEMENTAL DECLARATION OF SUSAN HICKEY

I, Susan Hickey, do hereby declare under penalties of perjury, pursuant to 28 U.S.C. §1746, that the following is true and correct to the best of my knowledge, information and belief:

1. I am the Supervisory Program Analyst, National Data Systems (NDS), Austin, Texas. NDS is a division of the Veterans Health Administration ("VHA") Office of Health Information of the Department of Veterans Affairs ("VA"). I previously submitted a declaration in this matter providing details about the searches that I supervised to locate responsive documents in response to the Freedom of Information Act (FOIA) requests involved in this litigation, FOIA requests numbered 13-03333-F and 13-03470-F. I understand that recently the Plaintiff has challenged aspects of VA's case, such as the methodology/ease with which I was able to accomplish re-identification of data.

2. As described in my previous declaration, I conducted searches of VHA records systems in response to requests 13-03333-F and 13-03470-F; in summary, I extracted the datasets responsive to the FOIA requests, applied the HIPAA Safe Harbor de-identification standard to the data, and stored the resulting datasets.

3. I studied the risk that the supposedly "de-identified" data discussed in para. 2 could be re-identified and I found it to be extremely easy to re-identify individual patients even after "de-identification" of the data using the HIPAA Safe Harbor method. There is a vast amount of literature online that describes re-identification and how to accomplish re-identification. (Note that VHA publishes the dataset metadata descriptions online at (http://www.virec.research.va.gov/RUGs/RUGs-Index.htm). Anyone with internet access, therefore, can learn what each field means within the VA datasets. Once datasets are obtained, therefore, a simple internet search will define the fields that are part of the datasets.

4. To re-identify a known patient, I matched known quasi identifiers, such as in the case I used in my prior declaration regarding Aaron Alexis, the Navy Yard shooter (I chose this individual because he is deceased). Below is a path using only six variables to result in identification of this single patient. I used the FY 2013 enrollment dataset as requested in FOIA-13-03333-F, after application of the Safe Harbor de-identification method, in order to identify the individual, as follows:

```
where RACE_D = 'Black or African American'
and SEX_BEST = 'M' --Gender
and R_AGE = '33' –Redacted Date of Birth
```

and R_ZIP_ENRL = 'b3/b6-76100' --Redacted ZIP Code in which an enrolled veteran resides

and PREFAC_D = '549' -- The preferred facility is the health care facility identified on an enrollee's application for health benefits as the site where he or she prefers to receive health services.

and ENRLPRIO = '2' -- Enrollment Priority 2 is for Veterans with service-connected disabilities rated 30% or 40% disabling

Once one identifies the individual in this type of scenario, one can access the individual's medical records.

5. Knowing very few identifiers about an individual, such as those listed for U.S. Congressmen, military generals, or other public figures on Wikipedia or those included in individuals' Facebook pages, provides an individual seeking information the ability to easily re-identify individuals, leading to access to the individuals' records. I did this for numerous patients. Gender, age (group), race, purple heart, combat location, number of dependents and many more obvious quasi identifiers are not unique on their own, but are very unique when combined. That uniqueness makes it easier to identify the individual because there are fewer and fewer individuals with that combination of variables.

6. If an individual seeking information on a patient did not know the variables such as those cited above, there are other variables that could be used in the dataset, based on what the individual knows about the patient; for example, the individual seeking information might know that the patient would not be enrolled in FY2010 (EFY10 = '0') because he was on active duty, or that his disability rating is thirty present (SCPER = '30'). Those and other known variables would allow an individual to be positively identified, opening up access to the remainder of the patient's medical records.

7. I was also able to re-identify a patient when the patient was unknown. This is commonly used by data mining companies who are contracted by pharmaceutical companies (C. Christine Porter, *De-Identified Data and Third Party Data Mining: The Risk of Re-Identification of Personal Information*, 5 Shidler J.L. Com. & Tech. 3 (Sep. 23, 2008) paragraph 7, *at* http://www.lctjournal.washington.edu/Vol5/a03Porter.html.). The most common method to re-identify is the patient geoproxy attack, where the records from a single patient (as established by VAID number, which uniquely identifies every patient in the VA healthcare datasets) who has visited multiple providers ("ord_prov" and "pcp_dss" fields, for example, which reveal such information and remain in the PHA datasets requested after application of Safe Harbor) are used to determine the patient's zip code by combining the data with other available information. (Emam, Khaled El and Arbuckle, Luk, Anonymizing Health Data: Case Studies and Methods to Get You Started, First Edition, O'Reilly Media, 2013, pp 141-143). Once a patient's zip code is learned, it is easier to narrow down who the patient is and, by combining the zip code with other available information, it is easier to identify the patient. Our datasets, for example, include provider type/specialty/taxonomy, where the care was provided and where a prescription was filled. Patients see and use providers and pharmacies that are close to where they live. National Provider Identifier (NPI) is a HIPAA regulation that establishes one unique identifier for each healthcare provider or system, for each health plan. Consequently, VHA providers are in this registry, which is online and free to search. Provider specialty/taxonomy and business and practice location zip codes are included. You can try one of our doctors at https://npiregistry.cms.hhs.gov/NPPESRegistry/NPIRegistryHome.do, NPI is 1013038744. This type of information assists in narrowing down the identity of a patient to just one.

8. In order to find unique patients by location, disease, drug, or any other quasi identifier, I simply write a query against the dataset of interest, group by each of the variables and add the statement, having count(*) = 1. Then I simply used a free online database to match the patient, such as a voter registration database. By way of example, I was able to re-identify a patient in Nevada by following this method. In order to accomplish re-identification of a patient, such as I accomplished per the examples above, an individual needs to know no more than basic database querying to find unique patients; I am aware that this basic querying is taught in some Texas high schools. In addition, online literature shows that data mining is cheap and getting cheaper. When I was Googling about individuals, online companies popped up that would match addresses against entire datasets. I was also able to accomplish re-identification easily after the Safe Harbor de-identification, using freely available online tools; I was not allowed to purchase any online service in my re-identification efforts.

9. I was able to accomplish patient re-identification very easily after application of the Safe Harbor method; the results yield many more unique patients.

Executed this 30th day of January 2014.

Susan Hickey
Supervisory Program Analyst
National Data Systems
Department of Veterans Affairs